Article

Non-technical skills assessments in undergraduate medical education: A focused BEME systematic review: BEME Guide no. 54

Gordon, Morris, Farnan, Jeanne, Grafton-Clarke, Ciaran, Ahmed, Ridwaan, Gurbutt, Dawne, Mclachlan, John Charles and Daniel, Michelle

Gordon, Morris orcid iconORCID: 0000-0002-1216-5158, Farnan, Jeanne, Grafton-Clarke, Ciaran, Ahmed, Ridwaan, Gurbutt, Dawne, Mclachlan, John Charles orcid iconORCID: 0000-0001-5493-2645 and Daniel, Michelle (2019) Non-technical skills assessments in undergraduate medical education: A focused BEME systematic review: BEME Guide no. 54. Medical Teacher, 41 (7). pp. 732-745. ISSN 0142-159X

Non-Technical Skills Assessments in Undergraduate Medical Education: A Focussed BEME Systematic Review

Morris Gordon, MBChB, PhD, MMed, Jeanne Farnan, MD, MHPE, Ciaran Grafton-Clarke MBChB, Ridwaan Ahmed, Dawne Gurbutt, John McLachlan, Michelle Daniel, MD, MHPE

Morris Gordon (MG), MBChB, PHD, MMed is a professor of evidence synthesis and systematic review, University of Central Lancashire, Preston, UK. Orchid.org/0000-0002-1216-5158

Jeanne Farnan (JF), MD, MHPE is Assistant Dean Curricular Development and Evaluation, and an Associate Professor, Section of Hospital Medicine at the University of Chicago Pritzker School of Medicine in Chicago, Illinois. Orcid.org/0000-0002-1138-9416

Ciaran Grafton-Clarke (CGC) is a medical student at Liverpool Medical school UK and was a summer intern at University of Central Lancashire, UK.

Ridwaan Ahmed (RA), MD, is a General Practitioner in Lancashire UK, and lecturer in public health sciences at the University of Central Lancashire, UK.

Dawne Gurbutt (DG), PhD is Acting Director in the Centre for Excellence in Learning and Teaching at University of Central Lancashire, UK.

John McLachlan (JM) PHD, is professor of Medical Education in the School of Medicine, University of Central Lancashire, UK

Michelle Daniel (MD) MD, MHPE is Assistant Dean for Curriculum, and an Assistant Professor of Emergency Medicine and Learning Health Sciences at the University of Michigan Medical School in Ann Arbor, Michigan. Orcid.org/0000-0001-8961-7119

Contact information (corresponding author)

Morris Gordon, HA244, Harrington building, University of Central Lancashire, Preston, UK, Tel 01772 201 201, mgordon@uclan.ac.uk

Disclosure of funding

None to disclose.

Abstract

Background

Many medical schools have implemented curricula to teach non-technical skills, a personal set of complex social and cognitive skills which are grounded in human factors safety industries in and out of health. Consensus on how to assess these skills is lacking. This systematic review aimed to evaluate the evidence regarding non-technical skills assessments in undergraduate medical education, to describe the tools used, learning outcomes and the validity, reliability and psychometrics of the instruments. Given the discrete context, a focussed review model is being deployed.

Methods

Studies describing assessment methods as either the focus of the study or having non-technical skills assessment as an outcome measure of the research were considered. A standardized search of online databases was conducted and consensus reached on included studies. Data extraction, quality assessment and content analysis were conducted per Best Evidence in Medical Education guidelines.

Results

Nine papers met the inclusion criteria. Assessment methods broadly fell into three categories: simulated clinical scenarios, objective structured clinical examinations, and questionnaires or written assessments. Details of methodology were synthesised to support readers developing their own materials. Tools to assess non-technical skills were often developed locally, in

response to specific educational interventions, without reference to conceptual frameworks. Consequently, the tools were rarely validated, limiting dissemination and replication. The majority of studies achieved outcomes modifying knowledge and skills of participants. Two studies resulted in behavioural change and one resulted in change in practice.

Conclusions

There were clear themes in content and broad categories in methods of assessments employed, with the OSCE identified as most able to assess multiple related skills at once. The quality of this evidence was poor due to lack of theoretical underpinning, with most assessments not part of normal process, but rather produced as a specific outcome measure for a teaching based study. Data on validity, reliability and learning outcomes was not available so these questions cannot be addressed at this time. Whilst the current literature forms a good starting position for educators developing materials, there is a need for future work to address these weaknesses as such tools are required across health education.

Keywords: Undergraduate medical education, non-technical skills, assessment, patient safety, human factors

Introduction

Non-technical skills describe a set of social and cognitive abilities encompassing *situational awareness, risk assessment, clinical decision making, leadership, communication skills and teamwork* (Gordon et al 2015a; Baldwin et al 1999*). Situational awareness and risk assessment* involve perceiving, understanding and anticipating risks in a given environment (e.g. a physician recognizes that task-switching during a busy shift contributes to errors, thus the physician deliberately slows down and focuses on only one task when reading an ECG). *Decision making* requires the ability to diagnose situations and make judgements concerning an appropriate course of action (e.g. a nurse suspects a haemolytic reaction and immediately stops the blood transfusion when his patient develops a fever and hypotension*). Leadership* describes an ability to influence others and provide direction without imposed hierarchies (e.g. a surgeon guides the operating team through a complex case while creating an environment that encourages all members of the team to speak up to prevent error). *Communication describes the key skills needed to share information across power and professional boundaries, ensure clear messages are produced and using techniques to ensure understanding and teamwork* describe unity around shared goals, defined roles and clear information exchange.

Many may equate non-technical skills with the field of human factors. Whilst Human Factors Ergonomics seeks to engineer entire systems to reduce errors, it recognises that non-technical skills are a specific area which can be fostered in individuals to support safety. Indeed, as the whole field of Human Factors is situated in the context of enhancing safety within the many industries in which they were developed, the medical and medical education literature

inescapably links the two. Lessons from other high reliability industries, such as aviation and the military, have been enthusiastically embraced within healthcare. For example, the discipline of human factors, environmental, organizational and job factors, and human and individual characteristics which influence behaviour (Wet, 2017), has been applied in aviation and the military to enhance the design of equipment, optmize the working environment and maximize performance (Catchpole, 2013). This educational methodology has been adapted for use successfully in fields such as anethesiology and shown to reduce error (Flin and Patey, 2009).

It is interesting to ponder whether such skills should form the basis of good medical practice across the board, rather than just error avoidance. However, currently almost all works in medicine that discuss non-technical skills do so in a safety context (Gordon et al., 2012, 2015b) and as the only skills that can be cultivated in the individual to enhance to achieve such goals, this link is not surprising. What is surprising is that although two-thirds of United States (US) medical schools mention patient safety in coursework (Blumenthal, 2010), only 25% describe curricula with explicit attention to safety skills-based training (Alper et al., 2009), but mention of non-technical skills conspicuous by its absence.

The World Health Organisation (WHO) curriculum on patient safety for medical students (2009) aims to encourage and facilitate the teaching of skills-based patient safety topics to medical students, with a specific focus on team-based training, systems and error prevention (Walton et al., 2010), but again there is no explicit mention of non-technical skills. There are some indications that this deficit is being tackled. More recently, organizations such as the American

Medical Association are championing change in undergraduate medical education, with specific focus on teaching new content in health systems science, which now include mention of non-technical skills (*Creating the medical school of the future,* 2017).

Non-technical skills should be delivered as part of the undergraduate core curriculum before professional attitudes are fully formed (Flin and Patey, 2009). The failure to incorporate such training into Undergraduate Medical Education UME may result in such topics being undervalued. Similarly, the absence of non-technical skills training in the postgraduate curriculum is a failed opportunity to provide repeated practice and to develop an integrated, longitudinal assessment strategy. Education and assessment of non-technical skills in UME learners may provide a pathway to achieving safer and more effective care.

.

A prior systematic review investigating non-technical skills educational interventions found that most were deployed with the expressed purpose to enhance patient safety. With regards to specific interventions, there was an identified a lack of scientific and theoretical rigor underpinning published teaching innovations (Gordon et al 2012), however, synthesis allowed for some existing theoretical constructs to be identified. A psychological theory of egocentric heuristics (Chang et al., 2010) describes a tendency to overestimate how well communication has been understood. Agency theory (Cheung et al., 2010) describes how in shift based working a focus on task rather than individual patients can only be challenged by error wisdom. Finally, the theory of 'coordination costs', and theories concerning the diffusion of responsibility (Darley and Latané, 1968) describe the role of systems, processes and technology in

counteracting such problems as systems grow and become more complex. Further work led to the development of the SECTORS model of non-technical skills learning (Systems and technology use, Error awareness, Communication, Team-working, Observation and simulation and Risk assessment and Situational awareness) (Gordon2013; Gordon et al., 2015a).

Educational interventions to enhance non-technical skills are necessary, but in isolation are not sufficient to advance the field within education or clinical practice. Such education must be coupled with rigorous assessments to both drive learning, and ensure competence. Unfortunately, non-technical skills are difficult to assess as they form part of a wider set of interconnected behaviours. Measuring the impact on patient outcomes also necessitates finding a way to assess the longitudinal impact of education.

A scoping review failed to reveal any systematic reviews investigating the assessment of non-technical skills within UME. Given the summative nature of many learning outcomes in the field, we feel such a review is vital to guide the inclusion of such assessments in high stakes summative examinations and to identify how they may be assessed in a methodologically robust manner. As such, we set out to systematically review the evidence regarding non-technical skills-based assessments in UME, to describe the overarching strategies utilized, learning outcomes addressed and the impact of these assessments, in terms of their validity, reliability, effect on performance and solutions to psychometric challenges. A focussed review methodology has been used, defined as ‘a form of knowledge synthesis in which the components of the systematic process are applied to facilitate the analysis of a more focused

research question' (Gordon et al., Under review). The focused review still embraces the core principles of systematic methodology, as these are crucial to faciliate transparency and scholarly deployement. However, after scoping the project and identifying the close link to patient safety of such skills, it became clear that the research scope was narrow and suited this mehtodology.

Methods

No single research paradigm was used for this review. We embraced both positivism, through description and justification of the assessment methods used, and constructivism, through clarification of the underpinning theoretical frameworks that informed assessment choice (Gordon, 2016). The manuscript was reported in accordance with the STORIES statement, publication standards for healthcare education evidence synthesis (Gordon and Gibbs, 2014) and the focussed review deployed in line with specific guidance (Gordon et al., under review).

Data collection

Scoping searches were performed to refine the search syntaxes and to clarify the inclusion and exclusion criteria for the review. We encountered two key problems during scoping: First, few papers described an assessment of non-technical skills as their primary focus. Second, few papers described an assessment tool at all. Thus, we broadened our initial searches and considered papers for inclusion if they either described an assessment as the main focus, or if they described an assessment as an outcome measure of an educational activity seeking to improve non-technical skills.

We embraced all study designs that targeted medical students, including when medical students participated within multidisciplinary teams and when the assessments formed a core or elective component of an undergraduate medical curriculum. Papers that described outcomes at all levels of Kirkpatrick's adapted hierarchy were eligible for inclusion (Gibbs, 2004). Studies from any country, published in any language were considered. Studies describing

only teaching without an assessment, failed to describe outcomes or described outcomes related to teaching, but not assessment, and papers that gave opinions or reviews without the primary use of an assessment tool were excluded. We excluded studies focusing on the assessment of non-technical skills in the post-graduate populations because this landscape has already moved in many ways to integrate formative assessment of such skills in the field of simulation. We believed that the specific needs within the undergraduate landscape, in particular summative assessment of such skills, represented a distinct educational problem and context for this review. (The search syntaxes, example search strategy, inclusion and exclusion criteria are summarized in Appendix 1 and 2).

The following online databases were searched from inception date of database up to January 2017 using a standardized search strategy: ERIC, PubMed, MEDLINE, EMBASE, CINAHL, Psychinfo, and Google Scholar. Abstracts available online from relevant education societies, including the Association for Medical Education in Europe (AMEE) and the Association for the Study of Medical Education (ASME) were also searched for the last 3 meetings to ensure any papers currently under review, but not fully published were included. Reference lists of the included studies were hand-searched for additional relevant studies.

Data analysis

Citations were screened independently by two authors, MG and RA. Abstracts considered potentially relevant for inclusion were independently reviewed by these same authors. Agreement was assessed using Cohen's kappa statistic. Full-text articles were then reviewed to

determine whether all inclusion criteria were met. Any disputes at any stage of the data analysis process were resolved by consensus. When insufficient information on an assessment was provided to make a judgment, we attempted to contact the authors for further details. For the included studies, the full manuscripts were assessed independently using a data extraction form (Appendix 3) by CGC and RA, with MG and DG ratifying the assessments.

The data extraction form and quality assessment tool (Appendix 3) were produced utilizing guidance from Best Evidence Medical Education (BEME) (Hammick et al, 2010), PRISMA (2017), and Reed et al. (2005).

The quality assessment of the included studies was broadly split into two main components: research methodology quality and reporting quality (Appendix 3). The research methodology quality assessment was completed as a 'yes/no' response to eight questions focusing on study objectives, study design, randomisation, reporting of participant characteristics and description of the intervention. The reporting quality assessment included six items: description of underpinning theoretical models, description of the assessment process, the educational context, psychometric details, provision of materials allowing replication and the strength of the conclusions drawn. The first five of these items were scored on a three-point Likert scale, with the last item *strength of conclusions*, scored against a five-point Likert scale. The impacts of the interventions were classified in accordance with Kirkpatrick's adapted hierarchy (Bates, 2004), in line with guidance provided by BEME (Hammick et al, 2010).**Error! Bookmark not defined.** A

descriptive synthesis of all included studies was completed, summarizing key findings, with an assessment of quality indicators as listed above.

If data was provided that supported quantitative analysis, such as validity or reliability, this was completed using the Cochrane Revman software (2014). For continuous data, the standardized mean differences (SMD) were compared. For discrete data, odds ratios (OR) were used. For data regarding the theoretical underpinning, pedagogy and content of the assessments, a posteriori thematic analysis (Strauss 1998) was planned in detail in the protocol, but as such data was ultimately not available, these details are not described.

Ethical approval was not sought for this review as it did not involve any direct participants.

Results

Search Results

Initial searching of both databases and alternative sources yielded 12,180 records, leaving 10,060 citations after de-duplication. After title screening, an additional 9463 citations were removed, leaving 597 abstracts to screen for eligibility. All abstracts were read by MG and RA, and 19 articles met the criteria for full-text assessment. The most frequent rationales for excluding studies at this stage included no assessment measure, a review article, letter or editorial, or an exclusive focus on graduate or post-graduate level learners. Given the relative clarity of such judgements, these were clear from the abstract with no full text review needed for most papers to exclude at this stage. No studies were excluded on the grounds of publication in a non-English language. Agreement between the two reviewers on abstract screening was good (kappa statistic: 0.91).

Of the 19 articles undergoing independent full-text assessment for inclusion, ten were excluded on the grounds of not meeting the full inclusion criteria (Anderson et al., 2009; Dudas et al., 2011; Hall et al., 2010; Kiesewetter and Fischer, 2015; Leung and Patil, 2010; Martinou et al., 2015; Meier et al., 2012; Myung et al., 2012; Robertson et al., 2010; Stahl et al., 2011). The reasons for all of the exclusions at this stage were related to not describing a non-technical skills assessment, but instead a limited outcome measure focussed on verifying the education delivered within these primary studies. The key discriminating factor used to make the decision was whether the assessment had any potential utility as part of ongoing formative or summative assessment, outside of the report itself. Given the difficulty of such judgements,

particularly as many of the included studies were similar, these were discussed amongst the whole team and consensus confirmed.

Nine papers were ultimately included in the qualitative and quantitative syntheses (Farnan et al., 2016; Jansson et al., 2015; Ginsburg et al., 2014; Daud-Gollotti et al., 2011; Madigosky et al., 2006; Aboumatar et al., 2012; Müller et al., 2012; Paxton and Rubinfeld, 2010; Thomas et al., 2015). The search flowchart is shown in Figure 1 (2006-16) and an overview of the included papers is shown in Table 1. Data was extracted independently by CGC and RA, who achieved concordance in 94% of quality ratings, with consensus reached on discussion (Appendix 4).

Study Participants

Of the nine studies, five of them were based in the United States (US) (Farnan et al., 2016; Jansson et al., 2015; Madigosky et al., 2006; Aboumatar et al., 2006; Paxton and Rubinfeld, 2010), with a further one in each of the following; Canada (Ginsberg et al., 2014),Brazil (Daud-Gallotti et al., 2011), Germany (Müller et al., 2012), and the United Kingdom (Thomas et al., 2015). The average number of participants per study was 92 (range: 18 – 214). All but three studies focused entirely on medical students (Farnan et al., 2015; Ginsberg et al., 2014; Paxton and Rubinfeld, 2010), with two studies covering second year medical students (Madigosky et al., 2006; Aboumatar et al., 2012), four studies focusing on third and fourth year medical students (Farnan et al., 2016; Jansson et al., 2015; Ginsberg et al., 2014), and the remaining three studies including fifth or sixth year medical students (Daud-Gollotti et al., 2011; Müller et al., 2012; Thomas et al., 2015). Two studies, in addition to undergraduate medical students,

involved nursing students (Ginsberg et al., 2014) or physician associate students (Paxton and Rubinfeld, 2010). There was minimal commonality between studies focused on participants at similar stages in medical education training, or between studies in the same country.

Quality Assessment

From a methodological perspective, all studies bar one included a review of the literature (Jansson et al., 2015) and provided clearly defined objectives (Paxton and Rubinfeld, 2010). All nine studies reported on and designed their study appropriately in response to the research question and provided learner characteristics. Three studies employed the use of control groups (Müller et al., 2012; Paxton and Rubinfeld, 2010; Thomas et al., 2015), three utilised a form of randomisation (Farnan et al., 2016; Aboumatar et al., 2012; Müller et al., 2012), and two studies described the educational intervention in enough detail as to allow for replication (Farnan et al., 2016; Daud-Gallotti et al., 2011) (Appendix 4).

Four (Farnan et al., 2016; Jansson et al., 2015; Madigosky et al., 2006; Aboumatar et al., 2012) of the nine papers did not provide any descriptions of theoretical models or conceptual frameworks utilized for the non-technical skills assessments. A further three provided limited descriptions aligned with conceptual elements related to error wisdom and situational awareness (Ginsburg et al., 2014; Müller et al., 2012; Paxton and Rubinfeld, 2010). The remaining two gave significant detail of the frameworks used, one fully aligning with theoretical principles (Daud-Gallotti et al., 2011) of non-technical skills (Ginsburg et al., 2014) and the other

identifying key theories related to error origins, describing the 'practice-effect' concept (Thomas et al., 2015).

Only two of the nine studies provided details of materials used, such as mark sheets, in sufficient detail as to allow replication (Müller et al., 2012; Thomas et al., 2015). The strength of conclusions estimated by employing the BEME strength of findings scale (Hammick et al., 2010), revealed eight studies scoring 3/5, suggesting that their conclusions were most likely based on the results (Farnan et al., 2016; Jansson et al., 2015; Daud-Gallotti et al., 2012; Madigosky et al., 2006; Aboumatar et al., 2012; Müller et al., 2012; Paxton and Rubinfeld, 2010; Thomas et al., 2015). Only one study achieved a score of 4/5, suggesting the conclusions are clear and very likely to be true (Ginsburg et al., 2014). Over half of the studies (5 of 9), provided a clear description of the process and outcomes of the assessment (Farnan et al., 2016; Ginsburg et al., 2014; Daud-Gallotti et al., 2011; Madigosky et al., 2006; Thomas et al., 2015).

Assessment Tools

The wide variation of educational interventions and non-technical skills assessed meant that a diverse range of assessment methods were utilized. The specific non-technical skills assessed varied, which is understandable given the lack of explicit recognition of what constitutes such skills until recently that has been previously noted, but attempts were made to consider how such skills fit within the more recent published skill sets. Described skills included situational awareness, distraction management and managing risk (seen as a subset of risk assessment), teamwork and maintaining interprofessional relationships (which also included elements of

communication), hazard identification (situational awareness), system thinking (clinical decision making), humanistic behaviour, self-efficacy and workplace attitudes (all elements of leadership) (Table 2). The variation in interventions and skills assessed was from more than just nomenclature and represented a primary source of significant educational heterogeneity among the studies. Broadly, assessments fell into three categories - simulated clinical scenarios, Objective Structured Clinical Examinations (OSCEs), and questionnaires or written assessments (Figure 2). An overview of the methods used in each paper is depicted in Figure 2.

Three studies employed simulated clinical scenarios as an assessment method with a variety of outcome measures. These included the Situation Awareness Global Assessment Technique (SAGAT) questionnaire (Müller et al., 2012), the standardized error and distractor management checklist (Thomas et al., 2015), and the Team Emergency Assessment Measure (TEAM) (Jansson et al., 2015). (See Table 2 for details of simulated clinical scenario assessment methodologies.) These assessment modalities tended to assess a limited number of non-technical skills, which may limit their utility in UME. However, the tools did to have good validity evidence for assessing specific skills. Thomas et al. (2015) was able to demonstrate the ability of simulated clinical scenarios to assess key behaviours through a standardised checklist, with a statistically significant reduction in medical error rates in the intervention group in practice.

Three studies utilized OSCEs. One study assessed learners' identification of patient safety hazards in a patient safety room of horrors (Farnan et al., 2015). This study aligned with error awareness as an underpinning element of non-technical skills (Gordon, 2013). Another study

evaluated patient safety competencies on a 5-point global rating scale (Ginsburg et al., 2014) and a third study utilized a patient safety checklist (Daud-Gallotti et al., 2011). (See Table 3 for details of OSCE assessment methodologies.) OSCEs appeared better able to assess multiple non-technical skills compared to simulated clinical scenarios.

Three studies utilized questionnaires and written assessment methodologies with a variety of assessment outcomes and little overlap. These included a validated Systems-Thinking Scale (STS) (Aboumatar et al., 2012), self-efficacy Likert scales (Aboumatar et al., 2012), and multiple-choice medical error assessments (Paxton and Rubinfeld, 2010). (See Table 4 for details of questionnaires and written assessments). All competency performances improved immediately after assessment when compared to the pre-intervention score across all three studies. Paxton et al. (2010) and Madigosky et al. (2006) performed long-term post-intervention testing in addition to immediate pre-and post-intervention testing, with only Paxton et al. demonstrating a statistically significant improvement in medical error competence after one year, when compared to the control group. Madigosky et al. (2006) reported that a number of competency scores changed in an undesired direction one year after the intervention.

Learning Outcomes

The level of impact for assessments used in the majority of included studies (6 of 9), sat at Level 2b of Kirkpatrick's hierarchy, which correlates to the modification of knowledge and skills (Farnan et al., 2016; Jansson et al., 2015; Ginsburg et al., 2014; Aboumatar et al., 2012l Paxton and Rubinfeld, 2010; Thomas et al., 2015). A further two studies impacted behaviour (Level 3)

(Madigosky et al., 2006; Müller et al., 2012), and one study impacted at the organization level, causing change in practice (Level 4a) (Daud-Gallotti et al., 2011). We could not complete a meta-analysis as there was significant assessment methodology heterogeneity and a lack of presentation of appropriate data

Validity and Reliability

No study determined the validity of the assessment tools used to assess non-technical skill performance, with the exception of the STS questionnaire (Aboumatar et al., 2012). As discussed previously, no quantitative analysis was performed relating to validity and reliability as no data was provided to allow this.

Discussion

This review identified a small number of studies that describe methods of assessing non-technical skills. As consensus on key elements that form non-technical skills has only been achieved in the past few years (Gordon and Gibbs, 2014) and even more recently such key elements adopted explicitly in any policy document on UME (*Creating the medical school of the future,* 2017), there was pervasive heterogeneity in the skills described that limits synthesis or even useful comparison. The review process was further hampered as the studies universally focused on reporting non-technical skills instruction as their primary goal, and assessment was viewed as a symbiotic component of this, allowing local programs to be verified or outcomes in practice to be tracked. Whilst included studies were only those judged to present assessment with the potential to form a continuing part of formative or summative assessment, the majority failed to provide the conceptual frameworks or theoretical models underpinning the choice of assessment. Psychometrics were almost completely absent and there were limited attempts to describe validity and reliability.

Assessment methods fell into three main categories: simulated clinical scenarios, OSCEs and questionnaire or written assessments. The methodological quality of the studies was of a reasonably good standard from a research perspective, but poor when considering the quality of reporting and evaluating the assessments themselves (Table 1). Despite many of the studies failing to evaluate the performance of the assessment method, a number of comparisons can be made across the different assessment modalities. First, studies utilizing simulated clinical scenarios as an assessment method were more likely to use validated and reliable assessment

tools. For example, the TEAM tool has a substantial body of normative data confirming its validity (Cooper and Cant, 2014). However, it only assesses a small component of the non-technical skill spectrum, and thus would only be called upon as an assessment measure of teamwork behaviours. The operational requirements to implement simulated clinical scenarios are significant and should be a major consideration when developing non-technical skill curricula and assessments, but similar to other areas of UME, assessment tools must be fit for purpose.

Our review leads us to conclude that there is a limited pool of non-technical skills assessment tools with an evidence-base supporting their use. The implications of this are that educators are increasingly dependent on developing their own assessment tools, with very little opportunity for reliability and validity testing. A single study demonstrated the reliability of assessing non-technical skill performance through an OSCE, suggesting a single assessor per station is sufficient, but generalization from this study is difficult. Given the unifying need for outcomes in non-technical skills to be met by learners (*Creating the Medical School of the Future*, 2017; Flin and Patey, 2009), not having similar evidence-based assessments is a major barrier to the field moving forward. There is not sufficient evidence to even suggest a specific model or conceptual framework to underpin assessment methods as given this lack of evidence, we would propose alignment with models that guide teaching of non-technical skills may be a best estimate (Gordon, 2013).

In considering these findings, readers must also consider some key limitations. First, while we aligned with a consensus definition of non-technical skills (Gordon et al., 2015), confusion persists around these terms and how they are applied. The search went through several scoping revisions and the kappa statistic demonstrates that within the team a clear consensus understanding was reached. However, the lack of consensus in the wider literature may have led to the author team's definition not fully matching the wider research or teaching body. Second, many of the studies primarily sought to report an intervention, with assessment devised to produce outcome measures. The inclusion of such papers is innately of lower quality from an assessment perspective. This is because researchers wishing to demonstrate effectiveness of teaching using a framework such as Kirkpatrick's hierarchy are not required to consider key issues for those designing formative and summative assessments for ongoing use (such as reliability, validity, particularly in the high stakes setting, cost and practicality, external scrutiny from university bodies and regulators). While we feel these are relevant when interpreted in that context, this does limit the utility of the findings and those wishing to employ them in a formal manner to formatively and summatively assess some outcomes would require many more aspects of the evidence to be presented to achieve this. Finally, as the goal of such outcomes is to enhance safety for patients, it is worth noting that all but one study failed to investigate patient outcomes (Daud-Gallotti et al., 2011). Subsequently, 'validity' has essentially not been demonstrated. 'Validity' would refer to how the assessments measure the non-technical skills in practice and their outcomes for patients.

While it is disappointing that such limited evidence has been identified, this review highlights that further work in this area is vital. First, future work must ensure that assessments align with truly ensuring outcomes that matter for care. This will allow validity in all its forms to be established, but this is challenging. Assessing the impact of education on practice is elusive, requiring as it does, the investigation of outcomes in clinical settings affected by multiple variables. This challenge is relatively common in health and medical education where the changes influencing patient outcomes may be cognitive and behavioural. It is crucial that assessment is sufficiently agile and robust to identify measurable elements which impact patients.

Current assessment modes traditionally favour models assessed close to training source and existing knowledge on more longitudinal and nuanced assessment of impact is limited. However, the nature and background of non-technical skills training provides the setting and stimulus to develop assessment modes which address these complex, problematic areas.

Second, quality elements of assessment research must be considered in detail, including clear mapping across local, national and international outcomes and psychometrics once used in practice. Educators may use the descriptive synthesis included in this review to support their current decision making, but assessments themselves must be the focus of future work, rather than a secondary component of research to implement non-technical skills education. This is key to ensure assessment has utility for both summative and formative assessment as part of ongoing UME.

Finally, consideration of the postgraduate setting is needed. This may be helpful first through a similar review piece as our scoping found no such review and then possibly through related future research. In achieving this, the issue of nomenclature is key. Using consensus terms for key skills in a consistent manner aids replication and dissemination and is key moving forward.

Conclusion

Simulated clinical scenarios, OSCEs and questionnaires have all been employed as methods to assess non-technical skills in undergraduate medical education. A multitude of assessment tools are used within these, including checklists, global rating scales and multiple-choice assessments and it is worth noting that the OSCE was most able to address multiple skill assessment. As studies typically report non-technical skills instruction as the primary goal, often the assessment methods described are not grounded in conceptual frameworks and key theoretical models, with further lack of clarity related to the use of variable language to describe specific skills. Educators are currently still dependent on developing their own assessment tools. This represents a major barrier to the field going forward, as such, there is a major requirement to develop and test evidence-based assessments of non-technical skills. There is an urgent need for research that focuses on developing non-technical skills assessments mapped to learning outcomes and consensus descriptions of specific skills.

They must be evaluated in terms of validity and reliability, addressing psychometric properties, as our review questions related to these areas cannot be adequately answered at present. All such research must be reported in a manner that supports dissemination to advance the field.

Commented [MP1]: *Morris please see if you agree with text below….The idea is to add something to justify why you only got 9 studies, This is important.*

_____

The authors were surprised by the very low number of accepted studies. This can be explained by the fact that studies typically report non-technical skills instruction as the primary goal, often the assessment methods described are not grounded in conceptual frameworks and, key theoretical models, with more clarity related to the use of variable language to describe specific skills are needed.

.

Commented [MP2]: *Morris please see if you agree ….The idea is to add something more than just to ask for more research*
Because research alone with no good primary studies will bring us nowhere.

_____

Moreover, there is an urgent need for primary studies of quality to address the assessment of non-technical skills. Educators are aware of the importance of assessment to drive learning and facilitate curricular changes (*if you want to change your curriculum, change your assessment*). This is why, Editors are asked to request the authors of manuscripts on non-technical skills to describe the assessment formats, their quality as well as students' results. This is essential to support teachers' decisions on their daily practice.

Appendices

Appendix 1: Inclusion and exclusion criteria

Criteria type	Inclusion criteria	Exclusion criteria
Study design	Any study design that describes the primary use of a method relating to the assessment of non-technical skills. Such assessments must have the potential for continuing use in UME outside of the specific context of the study.	Opinion pieces, editorial letters, commentaries, review articles which fail to describe the primary use of an assessment method relating to non-technical skills. They may also describe a limited assessment that is essentially a primary outcome measure and could not be used on a continuing basis in UME.
Outcome	Paper describes either an assessment tool for non-technical skills as the main focus of the study or in detail as an outcome measure of an educational activity (intervention) within a paper.	Paper describes teaching only without an assessment; Paper gives opinion or review but does not describe the primary use of an assessment method Paper describes an assessment method that is focussed on verifying the effectiveness of teaching in the context of research, not as a tool for learners (for either formative or summative assessment).
Outcome Assessment	Assessment of outcomes / impact at any level of Kirkpatrick's hierarchy	No assessment of outcomes / impact, OR Kirkpatrick's outcome measures focused exclusively on teaching, rather than outcomes of an assessment
Participants	All study designs targeting medical students, either in isolation or as part of a multi-disciplinary team	Study does not involve medical students

Stage of training	Assessment forms an elective or core component of an undergraduate medical curriculum	Assessment involves post-graduate (resident or continuing medical education) activities
Language	Any country, any language, with translation if needed.	

Appendix 2: Search syntaxes and example search strategy

Search Syntaxes:

(Non-technical skills OR human factors OR safety-training OR simulation) AND (medical student OR doctor) AND (health professions education course OR teaching OR training OR assessment OR medical education)

Stage	Adjoining word	Search term	Field to search	Number of results
1		Non-technical skills OR human factors OR safety-training OR simulation	Title	89,6297
2	AND	Non-technical skills OR human factors OR safety-training OR simulation	Title	
		Medical student OR doctor	Title	12,950
3	AND	Non-technical skills OR human factors OR safety-training OR simulation	Title	
		Medical student OR doctor	Title	
	AND	health professions education course OR teaching OR training OR assessment OR medical education	Title	7,556

Search strategy for the PubMed database

Appendix 3 – Data extraction form and quality assessment tool

Reference Number: *Reviewer*:

Source

☐Book	☐Comment	☐Conf.	☐Editorial
☐Guideline	☐Interview	☐Journal article	☐Lecture
☐Letter	☐News	☐Non-peer review article	
☐Official publication	☐Report	☐Thesis	

Citation information

First Author:

Title:

Search method

☐Electronic search ☐Hand search ☐Grey literature

☐Recommendation

Background/ question / objective [research methodology quality indicator]

Has a review of the literature been described? ☐Yes ☐No

Is there a clearly defined and well described objective to the study? ☐Yes ☐No

Research design [research methodology quality indicator]

Is the design appropriate to answer the research question?

Is the study design reported?

Place an S for Stated or I for Implied in the box:

☐Audit	☐Action-based
☐Survey	☐Cross-sectional study
☐Case-series	☐Observational
☐Retrospective cohort study	☐Prospective cohort study
☐Before-and-after study	☐Time series
☐Randomised trial	☐Non-randomised trial

Was a control group used? ☐Yes ☐No

Was there any form of randomization between groups? ☐Yes ☐No

Were the learner characteristics reported? (If NO continue to intervention)

☐Yes ☐No

Which groups were studied? (Please tick all that apply if mixed)
☐Doctor ☐Midwife ☐Nurse ☐Other

Were the study participants undergraduate or postgraduate?
☐Undergraduate ☐Postgraduate

Number of participants:
Demographics of participants:

Intervention [research methodology quality indicator]
Is the educational intervention clearly described? ☐Yes ☐No
Is the educational intervention described in enough detail to replicate? ☐Yes ☐No

Please record details of the intervention/assessment:

Is there a description of theoretical models or conceptual frameworks that underpin the choice of assessment? **[Underpinning framework quality indicator]**
Yes Clear and relevant description
Yes Some limited discussion of underpinning, with minimal interpretation in the context of the assessment choice
No

Is there a description of the process and outcomes of the assessment? **[Assessment method quality indicator]**
Yes Clear description of the process and outcomes of the assessment
Yes Some limited description that will not facilitate replication
No

Are details of the educational context and learner characteristics of the study provided? **[Background quality indicator]**
Yes Clear details of the educational context and learner characteristics of the study
Yes Some description, but not significant as to support dissemination
No

Are there details of psychometrics and how they are applied to the assessment? **[Psychometrics quality indicator]**

Yes Clear description of relevant psychometrics and how applied to this assessment
Yes Some psychometric information, but not enough to fully inform for dissemination
No

Is there provision of material to allow assessment replication? **[Content quality indicator]**
Yes Provision of detailed materials to allow assessment replication
Yes Some elements of materials presented or summary information
No

Results and strength of conclusions
What are the key conclusions?

Do the conclusions match the findings of the study? **[Strength of conclusions quality indicator]**
Low ⟶ High
1 2 3 4 5

1 – No clear conclusions can be drawn. Not significant
2 – Results ambiguous, but there appears to be a trend.
3 – Conclusions can probably be based on the results.
4 – Results are clear and very likely to be true.
5 – Results are unequivocal.

Did the research discuss limitations of the study?

Impact of intervention studied (target of evidence/ outcomes)

Do outcomes match the objectives of the study?
Are data collection methods described in enough detail to replicate? ☐Yes ☐No
Are statistical tests used? ☐Yes ☐No

If used, are statistical tests appropriate for the design?

Code the level of impact being studied in the item and summarize any results of the intervention at the appropriate level. Note: include both predetermined and unintended outcomes.
☐Level 1 - Participation
☐Level 2a - Modification of attitudes/perceptions
☐Level 2b - Modification of knowledge/skills

☐Level 3 - Behavioural change
☐Level 4a - Change in organizational practice
☐Level 4b - Benefits to patient / clients

Appendix 4 – Quality assessment of included studies

Author / Quality	Farnan	Jansson	Ginsberg	Gallotti	Madigosky	Aboumatar	Müller	Paxton	Thomas
Literature review described?	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Clear study objectives?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes
Appropriate study design?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Is the study design reported?	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Was a control group used?	No	No	No	No	No	No	Yes	Yes	Yes
Randomisation between groups?	Yes	No	No	No	No	Yes	Yes	No	No
Learner characteristics reported?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Educational intervention described?	Yes	No	No	Yes	No	No	No	No	No
Description of theoretical models?	No	No	Some	Clear	No	No	Some	Some	Clear
Description of the process?	Clear	Some	Clear	Clear	Clear	Some	Some	Some	Clear
Learner characteristics provided?	Some	Some	Some	Clear	Some	Clear	Some	Some	Some
Details of psychometrics?	Clear	No	No	Clear	Some	Some	Some	No	Some
Materials provided for replication?	Some	No	Some	Detailed	Some	No	Detailed	Some	Detailed
Conclusions match the findings?	3, 4	3, 3	4, 4	3, 3	3, 3	3, 3	3	3	3
Are study limitations discussed?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Outcomes patch	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Reproducible data collection?	Yes	No	Yes	Yes	Yes	No	Yes	No	Yes
Appropriate statistical tests?	N/A	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Level of Kirkpatrick's hierarchy?	2b, 2b	2b, 2b	2b, 2b	4a, 4a	3, 3	2b, 2b	3	2b	2b

Table 2 – Use of simulated clinical scenarios to assess non-technical skills

Assessment methodology	Non-technical skill assessed	Summary of assessment methodology	Example items	Use in non-technical skills assessment	Overview of scenario	Administration of assessment
Situation Awareness Global Assessment technique (SAGAT) (Muller et al., 2012)	Situation awareness	SAGAT is an objective measure of a students' situational awareness during a simulated scenario. Each item assesses situation awareness with a focus on perception, recognition or anticipation	What is the current oxygen saturation? (perception) What is the junior doctor doing? (recognition) How old is the patient? (perception) What is the number of white blood cells in the medical examination report? (perception) How will the blood pressure change in the next minute? (anticipation) How will the central venous pressure change in the next minute? (anticipation)	SAGAT questionnaire I and II were used at 4-6 minutes and 8-10 minutes respectively, during a 10 minute clinical scenario. Performance was assessed using the SAGAT during a pre- and post-intervention clinical scenario.	47-year old male patient was simulated who had just been transferred to ITU. The patient was in septic shock with suspicion of valve endocarditis one-year post-valve replacement (pre-intervention scenario). 29-year old female patient was simulated who was suffering from septic shock due to post-partum sepsis. (post-intervention scenario).	The scenarios were interrupted between 4-6 minutes and between 8-10 minutes. Participants were taken out of the simulated environment and had to complete a questionnaire with 10 items (first interruption) and 11 items (second interruption).
Standardised error and distractor management checklist. (Thomas et al., 2015)	Managing distraction and interruption	Dichotomous standardised checklist measuring the number of errors made, number of distractions and interruptions managed.	Did not correctly prioritise patients and order in which to be seen. Checks antibiotic vial as satisfactory when it has actually passed its expiry date. Does not check the identify of the relative wishing to speak to them gives information pertaining to a different patient. Identifies (and deals with) radio playing in the background. Identifies (and deals with) FY1 pager going off. Deal with request to prescribe Paracetamol for a separate unrelated patient.	Standardised checklist used to document student performance at both baseline and post-intervention (targeted feedback).	Participants acted as foundation doctors in 30-minute simulated ward round. Patient 1 had sepsis. Patient 2 had a post-operative myocardial events. Patient 3 was confused. Facilitators deployed six realistic, time-critical distractions and interruptions (radio, hoovers, additional prescription tasks, phone calls and upset relatives). Each patient had a number of expected tasks, with associated potential errors (e.g. wrong drug doses, incorrect CURB-65 calculation, incorrect patient identity).	1 month between the pre-intervention and post-intervention assessment (via the standardised checklist). Checklist completed by a trained member of the simulated ward round faculty.
Team Emergency Assessment Measure (TEAM) (Jansson et al., 2015)	Teamwork	TEAM scoring is an externally validated tool for assessing teamwork, under the domains of leadership, teamwork and task management. The 12 items (11 specific and 1	The team leader let the team know what was expected of them through direction and command (leadership). The team communicated effectively (teamwork). The team acted with composure and control (teamwork).	TEAM assessment tool was used to assess student performance following high-fidelity simulated resuscitation scenarios.	Simulated-based cardiac dysrhythmia session. No further information available.	Trained observers completed the TEAM assessment tool after viewing live or videotaped sessions.

		global rating) are rated using a five-point scale.	The team prioritised (task-management) The team followed approved standards and guidelines (task-management)			

Table 3 – Use of Objective Structured Clinical Examinations (OSCEs) to assess non-technical skills

Assessment methodology	Non-technical skill assessed	Summary of assessment methodology	Example items	Use in non-technical skills assessment	Overview of scenario	Administration of assessment
Identification of patient safety hazards (Farnan et al., 2016)	Patient safety hazard identification	9 simulated patient safety hazards. Items developed based upon discussion amongst local leaders in patient safety and strategic priorities for local hospital environments.	Inappropriate catheterisation. Mislabelled medication. Restraints in place. No prophylaxis written in medication list. No personal protective equipment provided in a patient with C. Diff infection. Patient with a clinical history of delirium, inappropriate bed height and catheter generating fall risk.	Participants faced the patient safety room of horrors. Participants were asked to identify safety hazards following the 10-min exercise. The participants were given 5 minutes after the simulation to complete the assessment task.	'Room of Horrors' clinical scenario which highlighted specific safety hazards and a mock patient chart were written, including clinical information about case, allergy status, medication list and a mock sign-out for the learner.	Participants given 5 min after the simulation to complete the online form documenting hazards identified. Students given a clipboard to note the hazards as they proceeded through the simulation. Students used Bline (simulation and clinical skills video and evaluation software) or ScanTron sheets to record hazards.
5-point global rating scales on patient safety competencies (Ginsburg et al., 2014)	Patient safety competencies (communication, teamwork, managing risk, disclosure and culture)	OSCE station scored on four or five patient safety competency dimensions (communication, teamwork, managing risk, disclosure and culture). A single 5-point global rating scale was developed.	No information available.	Participants were scored using the rating scales during a four-station OSCE, focused on the socio-cultural aspects of patient safety.	Station 1 required learners to uncover a deep vein thrombosis near miss and explain the factors to a spouse. Station 2 involved team dynamics and communication with a patient around a complex discharge. Station 3 required learners to persist in an interaction with a dismissive, time-pressured staff physician. Station 4 required learners to discuss an insulin overdose with the patient, including how it happened. Stations 8 min long.	Scoring in each station was done by two assessors who were both in the exam room. Each assessor was a current or retired faculty member from nursing or medicine. All assessors participated in a one-hour training session prior to the OSCE.
Patient safety checklist (Daud-Gallotti et al., 2011)	Medical error, interprofessional relationship and humanistic behaviour	Checklist contains 21-items in 3 domains (medical error, patient-physician relationship, humanistic behaviour). Competency in the medical error and patient-physician relationship domains were rated as follows: non-existent = 0 points, present but insufficient = 50 points and present & adequate = 100 points. Competency in the humanistic behaviour domain was evaluated with 5-point Likert scales.	Did the student tell you that a preventable adverse event occurred during your admission? (medical error recognition); Did the student identify preventative actions to avoid this error? (medical error prevention); Did the student introduce self to you before the interview? (verbal communication); Did the student respect your silence? (patient-centred care); Did you feel supported in your distress? (support); Did the student respect your rights and values? (respect)	Assessor scored the students using the patient safety checklist during each of the five OSCE stations.	Station 1: A 70-yo man with renal insufficiency and lumbar pain received an anti-inflammatory prescription during his hospitalisation. His renal function progressively deteriorated, and on day 4, dialysis was indicated. Task: explain to the daughter what happened to her father; Station 2: A 50-yo female with pneumonia was admitted to the ward overnight. In the same room was another person with a similar name who also had diabetes. On the following day, Patient 3 received insulin instead of Patient 2 and presented with confusion due to hypoglycaemia. She received glucose and recovered completely.	The standardised patients completed one checklist for each student during a 3-minute break between the exit of the previous student and the entrance of the next student.

Table 4 – Use of questionnaires and written assessments to assess non-technical skills

Assessment methodology	Non-technical skill assessed	Summary of assessment methodology	Example items	Use in non-technical skills assessment	Administration of assessment
System-thinking scale (STS) (Aboumatar et al., 2012)	System thinking	Externally validated scale used to measure system thinking. It is composed on 20 items on a 0 – 4 Likert-type scale. The composite score ranges from 0 to 80.	Item 2: 'I look beyond a specific event to determine the cause of the problem'. Item 10: 'I propose solutions that affect the work environment, not specific individuals' Item 14: 'I think small changes can produce important results'.	Used to assess system thinking before and after a six-hour patient safety curriculum.	Online
Self-efficacy Likert-type scale (Aboumatar et al., 2012)	Self-efficacy	Nine 'I know how to' statements on a 1-5 Likert-type scale.	'I know how to disclose a medical error' 'I know how to use personal protective equipment such as gowns, gloves and masks' 'I know how to use teach back'	Used to assess self-efficacy before and after a six-hour patient safety curriculum.	Online
Safety knowledge assessment (Aboumatar et al., 2012)	Safety	19-item safety test.	No data available.	Used to assess safety knowledge before and after a six-hour patient safety curriculum.	Online
Agreement statements (Aboumatar et al., 2012)	Intention to apply safety practices	Intention to apply safety practices is assessed via agreement ratings to two statements.	Statement 1: 'I will speak up about any safety concerns I have about my patients' Statement 2: I plan to use the Teach Back method to ensure that my patients understood my instructions'.	Used to assess the intention to apply safety practices following before and after a six-hour patient safety curriculum.	Online
Medical error knowledge assessment (Paxton and Rubinfeld., 2010)	Medical error knowledge	12-question knowledge assessment in a multiple-choice format. Questions were under the domains of: terminology, active versus latent error, incidence, error theory, error disclosure and legal considerations.	Question 1: 'define adverse event' Question 10: 'define malpractice' Question 12: 'identify high-risk specialities'	Used to assess medical error knowledge before and after a two-hour teaching intervention. Additionally, the assessment was used to assess long-term outcomes (up to 12 months)	Online
28-item questionnaire (Madigosky et al., 2006)	Knowledge, skills and attitudes	28-item questionnaire. 5 items assessed knowledge, 5 items measured skills, and 18 items assessed attitudes. All items were measured on a 5-point ordinal scale.	Attitude item: 'Competent physicians do not make medical errors that lead to patient harm' Skill item: 'Supporting and advising a peer who must decide how to respond to an error' Knowledge item: no data.	Used to assess knowledge, skill and attitudes before and after a 10.5-hour patient safety curriculum. Additionally, the assessment was used to assess long-term outcomes (12 months)	Online

References

Aboumatar, H., Thompson, D. and W,u A. (2012) 'Development and evaluation of a 3-day patient safety curriculum to advance knowledge, self-efficacy and system thinking among medical students', *BMJ Qual Saf*, 21(5), pp. 416-422.

Alper, E., Rosenberg, E., O'Brien, K., Fischer, M., *et al*. (2009) 'Patient safety education at U.S. and Canadian medical schools: results from the 2006 Clerkship Directors in Internal Medicine survey', *Acad Med*, 84(12), pp. 1672-1676.

Anderson, E., Thorpe, L., Heney, D. and Petersen, S. (2009) 'Medical students benefit from learning about patient safety in an interprofessional team', *Med Educ*. 43(6), pp. 542-552.

Baldwin, P.J., Paisley, A.M. and Brown S.P. (1999) 'Consultant surgeons' opinion of the skills required of basic surgical trainees', *Br J Surg,* 86(8), pp.1078-1082.

Bates, R. (2004) 'A critical analysis of evaluation practice: The Kirkpatrick model and the principle of beneficence', *Eval Program Plann*, 27(3), pp. 341-347.

Blumenthal, D. (2010) '*Patient Safety: Conversation to Curriculum'*, *New York Times*. Available at: http://www.nytimes.com/2010/01/26/health/26error.html?_r=1. (Accessed 9 October 2017)

Catchpole, K. (2013) 'Spreading human factors expertise in healthcare: untangling the knots in people and systems', *BMJ Saf Qual,* 22(10), pp. 802-808.

Chang, V.Y., Arora, V.M., Lev-Ari. S., D'Arcy, M. *et al*. (2010) 'Interns overestimate the effectiveness of their hand-off communication', *Pediatrics*, 125(3), pp.491-496.

Cheung, D., Kelly, J. and Beach, C. (2010) 'Improving Handoffs in the Emergency Department', *Ann Emerg Med*, 55(2), pp. 171-180.

Cochrane Collaboration 2014. Review Manager (RevMan) [Computer program]. Version 5.3. Copenhagen: The Nordic Cochrane Centre

Cooper, S. and Cant, R. (2014) 'Measuring non-technical skills of medical emergency teams: An update on the validity and reliability of the team emergency assessment measure (TEAM)', *Resuscitation,* 85(1), pp. 31-33.

Creating the Medical School of the Future. AMA Accelerating Change in Medical Education. https://www.ama-assn.org/innovations-outcomes-consortium. Accessed October 9, 2017.

Darley, J. and Latané. B. (1968) 'Bystander intervention in emergencies: Diffusion of responsibility', *J Pers Soc Psychol*, 8(4), pp. 377-383.

Daud-Gallotti, R., Morinaga. C., Arlindo-Rodrigues. M., Velasco. I., *et al*. (2011) 'A new method for the assessment of patient safety competencies during a medical school clerkship using an objective structured clinical examination', *Clinics (Sao Paulo)*, 66(7), pp. 1209-1215.

Dudas, R., Bundy, D., Miller, M. and Barone, M. (2011) 'Can teaching medical students to investigate medication errors change their attitudes towards patient safety?', *Qual Saf Health Care*, 20(4), pp. 319-325.

Farnan, J., Gaffney, S., Poston, J. et al. (2016) 'Patient safety room of horrors: A novel method to assess medical students and entering residents' ability to identify hazards of hospitalisation', *BMJ Qual Saf,* 25(3), pp. 153-158.

Flin, R. and Patey, R. (2009) 'Improving patient safety through training in non-technical skills', *BMJ*, 339, pp. 985-986.

Ginsburg, L., Tregunno, D., Norton, P. et al. (2014) 'Development and testing of an objective structured clinical exam (OSCE) to assess socio-cultural dimensions of patient safety competency', *BMJ Qual Saf,* 24(3), pp. 1-7.

Gordon, M., Darbyshir,e D. and Baker, P. (2012) 'Non-technical skills training to enhance patient safety: A systematic review', *Med Educ*, 46(11), pp. 1042-1054.

Gordon M. (2013) Building a theoretically grounded model to support the design of effective non-technical skills training in healthcare: The SECTORS model. J Contemp Med Edu. 2013;1(2) pp. 77–82

Gordon, M. and Gibbs, T. (2014) 'STORIES statement: publication standards for healthcare education evidence synthesis', *BMC Medicine*, 12, pp. 143.

Gordon, M., Box, H., Farrell, M. and Stewart, A. (2015a) 'Non-technical skills learning in healthcare through simulation education: integrating the SECTORS learning model and complexity theory', *BMJ Simul Technol Enhanc Learn*, 1, pp. 67-70.

Gordon, M., Baker, P., Catchpole, K., Darbyshire, D. *et al*. (2015b) 'Devising a consensus definition and framework for non-technical skills in healthcare to support educational design: A modified Delphi study', *Med Teach*, 37(6), pp. 572-577.

Gordon, M. (2016) 'Are we talking the same paradigm? Considering methodological choices in health education systematic review', *Med Teach*, 38(7), pp. 746-750.

Gordon M, Grafton-Clarke C, Hill E, Gurbutt D, Patricio M, Daniel M. (Under review). Twelve tips for undertaking a focussed systematic review in medical education. Medical Teacher, Under review

Hall, L., Scott, S., Cox, K., Gosbee, J. *et al.* (2010) 'Effectiveness of patient safety training in equipping medical students to recognise safety hazards and propose robust interventions', *Qual Saf Health Care*, 19(1), pp. 3-8.

Hammick, M., Dornan, T. and Steinert, Y. (2010) 'Conducting a best evidence systematic review. Part 1: From idea to data coding. BEME Guide No. 13', *Medical Teacher*, 32, pp. 3-15.

Jansson, P., An-Grogan, Y., Susan, G. *et al.* (2015) 'A Needs Assessment in Patient Safety Education for Fourth-Year Medical Students', *Am J Med Qual*, 30(6), pp. 601.

Kiesewetter, J. and Fischer, M. (2015) 'The Teamwork Assessment Scale: A Novel Instrument to Assess Quality of Undergraduate Medical Students' Teamwork Using the Example of Simulation-based Ward-Rounds', *GMS J Med Educ*, 32(2), pp. 1-9.

Leung, G. and Patil, N. (2010) 'Patient safety in the undergraduate curriculum: medical students' perception', *Hong Kong Med J,* 16(2), p. 101-105.

Madigosky, W., Headrick, L., Nelson, K., Cox, K. *et al*. (2006) 'Changing and Sustaining Medical Students' Knowledge, Skills, and Attitudes about Patient Safety and Medical Fallibility', *Acad Med*, 81(1), pp. 94-101.

Martinou, E., Chindambaran, R., Krishnasamy, R., Johnson, A. *et al*. (2015) 'Simulation in undergraduate medical education designing a programme to improve medical students' non-technical skills', *Int J Surg,* 23:S102.

Meier, A., Boehler, M., McDowell, C., Schwind, C. *et al*. (2012) 'A surgical simulation curriculum for senior medical students based on TeamSTEPPS', *Arch Surg,* 147(8), pp. 761-766.

Müller, M., Hänsel, M., Winkelmann, A. *et al*. (2012) 'Impact of simulator training and crew resource management training on final-year medical students' performance in sepsis resuscitation: A randomized trial', *Minerva Anestesiol*, 78(8), pp. 901-909.

Myung, S., Shin, J., Kim, J., Roh, H. *et al*. (2012) 'The patient safety curriculum for undergraduate medical students as a first step toward improving patient safety', *J Surg Educ,* 69(5), pp. 659-664.

Paxton, J. and Rubinfeld, I. (2010) 'Medical errors education: A prospective study of a new educational tool', *Am J Med Qual*, 25(2), pp. 135-142.

PRISMA. (2015) *Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)*. Available at: http://prisma-statement.org/ (Accessed: 29 August 2017).

Reed, D., Price, E., Windish, D. *et al*. (2005) 'Challenges in systematic reviews of educational intervention studies', *Ann Intern Med,* 142, pp. 1080-1089.

Robertson, B., Kaplan, B., Atallah, H., Higgins, M. *et al*. (2010) 'The use of simulation and a modified TeamSTEPPS curriculum for medical and nursing student team training', *Simul Healthc*, 5(6), pp. 332-337.

Stahl, K., Augenstein, J., Schulman, C.I., Wilson, K. *et al*. (2011) 'Assessing the impact of teaching patient safety principles to medical students during surgical clerkships', *J Surg Res,* 170(1), pp. 29-40.

Strauss, A.L. and Corbin, J.M. (1998) *Basics of Qualitative Research: Procedures and Techniques for Generating Grounded Theory. 2nd ed*. Thousand Oaks, CA: Sage Publications.

Thomas, I., Nicol, L., Regan, L. *et al*. (2015) 'Driven to distraction: a prospective controlled study of a simulated ward round experience to improve patient safety teaching for medical students', *BMJ Qual Saf*, 24, pp. 154-161.

Walton, M., Woodward, H. and Staalduinen, S. (2010) 'The WHO patient safety curriculum guide for medical school', *Qual Saf Health Car,* 19, pp. 542-546.

Wet, C. De. (2012) *An overview of patient safety in primary care*, *NHS Educ Scotl*. Available at: http://www.nes.scot.nhs.uk/media/2075343/an-overview-of-patient-safety-in-primary-care-nov-12.pdf (Accessed: 9 October 2017).